12

Transaction Costs and Liquidity Risk

Understanding and managing transaction costs is a critical component of portfolio risk analysis. Optimal rebalancing and hedging policies are heavily affected by consideration of transaction costs. Also, liquidity risk, which is the uncertainty connected to the ability to liquidate or rebalance a portfolio at a "fair price," is a very important component of portfolio risk, particularly during periods of market turmoil.

Section 12.1 provides some basic definitions. Section 12.2 discusses theoretical and econometric models of transaction costs. Section 12.3 looks at the time-series behavior of transaction costs and liquidity and their correlation with market movements. Section 12.4 considers optimal trading strategies in the presence of transaction costs and liquidity risk.

12.1 Some Basic Terminology

Markets for trading assets can take various forms: from decentralized search and negotiation (as for houses and used cars) to centralized electronic exchanges. Each market has its own set of rules that determine acceptable order types, priorities for order execution, and other attributes. We do not attempt to discuss all of these features in detail but instead characterize the main features that are common to most organized financial markets.[1]

So far in this book we have treated each asset price as having a unique value at each point in time. In fact, there are several definitions of an asset price at any time t. A trader may post a *limit order*, or *quote*, which is an order either to purchase or to sell a certain amount of an asset at the best price available, subject to a limit on the price. The quote specifies the direction of the trade (*buy* or *sell*), the *limit price*, the *size* of the trade, the *length of time* the order should be open, and other features of the order. For example, a limit buy order for 500 shares of XYZ Inc. at a limit price of $50.00 per share executes, or gets *filled*, if

[1] See Harris (2003) for a more detailed treatment of market structures.

there is ample inventory of XYZ at \$50.00 per share or less. A *partial fill* occurs if there is some inventory of XYZ at \$50.00 per share but not enough to fill a 500-share order. Note that both fills and partial fills may involve multiple trades and prices. The limit price is called a *bid price* for an order to buy and an *ask price* (or *offer price*) for an order to sell.

The menu of outstanding limit orders is called the *limit order book*. Ordering the quoted prices from highest to lowest, we first see orders to sell (ask prices) followed by orders to buy at lower prices. The total number of shares for which standing orders exist at a given price is called the *depth* of the limit order book at that price.

The *best bid and offer* (BBO) quotes are the highest bid price, p_t^{b}, and lowest ask price, p_t^{a}, in the limit order book. The difference between the best prices is called the *bid–ask spread*:

$$s_t = p_t^{\mathrm{a}} - p_t^{\mathrm{b}}.$$

Trades that are larger than the quote depth induce a less favorable price (a lower bid or a higher ask); the change in price in response to an order larger than the quote depth is called the *price impact*. Assuming no price impact, a round-trip purchase and immediate sale of an asset costs the trader the bid–ask spread. For a one-way trade, the relevant cost is half the spread. The average of the BBO quotes is called the *midpoint price*:

$$\mathrm{mid}_t = \frac{p_t^{\mathrm{a}} + p_t^{\mathrm{b}}}{2}.$$

The *proportional spread* is given by

$$s_t^{\mathrm{prop}} = \frac{p_t^{\mathrm{a}} - p_t^{\mathrm{b}}}{\mathrm{mid}_t}. \tag{12.1}$$

The proportional spread is a measure of the spread cost on a relative basis. The price at which an actual trade takes place is called the *transaction price*, which is denoted by p_t. In the simplest case, in which the market maker or another liquidity provider takes one side of each order and each trade is executed at the best bid price p_t^{b} or the best ask price p_t^{a}, we can view the historical record of transaction prices as a random sequence of bid and ask prices. It is important to note that this interpretation holds only in the simplest case when there is no price impact and all trades execute at the bid or ask prices. We discuss more general interpretations of the transaction price record below.

A *market order* is a directive to trade immediately at the best available price. For example, assume that there are limit orders to buy 400 shares of XYZ at \$50.00 and 600 shares of XYZ at \$49.98 and to sell 700 shares

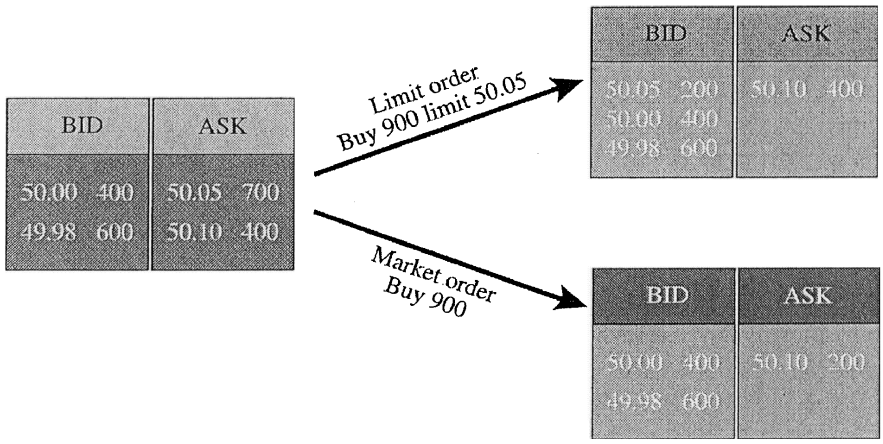

Figure 12.1. Changes in the limit order book due to different types of orders.

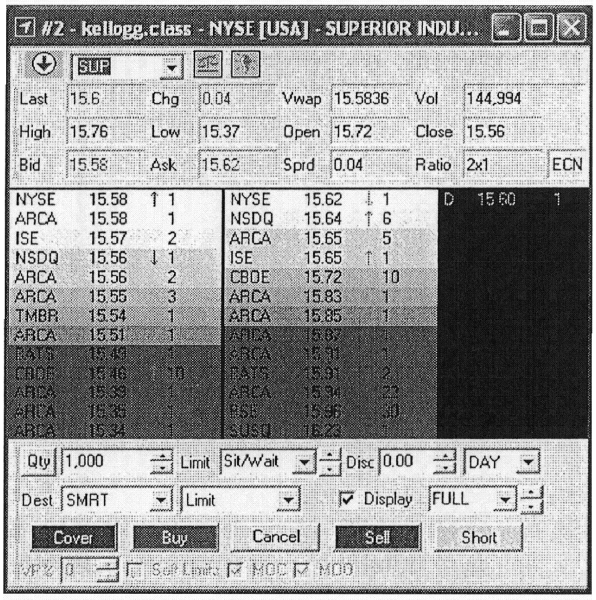

Figure 12.2. Limit order book for Superior Industries common stock.

of XYZ at \$50.05 and 400 shares at \$50.10, as shown in the left-hand side of figure 12.1. (There may be other limit orders further away from the best bid and offer quotes.) The quote midpoint is $\text{mid}_t = \$50.025$. A trader entering a market order to buy 900 shares would pay \$50.05 for the first 700 shares and \$50.10 for the next 200 shares, at an average price of \$50.0611, as illustrated in the lower right of figure 12.1. One measure of the cost of the trade is the difference between the transaction

price and mid_t. For this hypothetical market order, the cost is

$$p_t - \text{mid}_t = \$50.0611 - \$50.025 = \$0.0361,$$

which is equal to half of the spread, $p_t^a - \text{mid}_t = \$0.025$, plus the additional price impact, equal to $p_t - p_t^a = \$0.0111$.

Some markets do not allow market orders. Nevertheless, a trader can effectively submit a market order in the form of a *marketable limit order*: that is, a limit order that crosses the order book. In the example above, a limit order to buy 900 shares at \$50.10 is equivalent to a market order for 900 shares. However, a limit order to buy 900 shares at \$50.05 is equivalent to a 700 share market order plus a 200 share limit order at \$50.05: the first 700 shares would get filled by the standing limit order to sell at \$50.05 and the remaining 200 shares will not be executed (since the next limit order to sell is at \$50.10). The new best bid is \$50.05 rather than \$50.00, as illustrated in the upper right of figure 12.1. Figure 12.2 shows a snapshot of the limit order book for Superior Industries. Limit orders to buy are in the left panel, arranged from the best bid of \$15.58 at the top. Limit orders to sell are in the right panel, arranged from the best ask of \$15.62 at the top. The column to the right of the price column shows the number of shares for that limit order in units of 100 shares. The depth at the best bid is two round lots, or 200 shares, and the depth at the best ask is 100 shares.

The properties of limit and market orders tend to be complementary. For example, a standing limit order, which is one that does not execute immediately, supplies liquidity to the market since other traders have the option of taking the other side of the order at any time. In contrast, a market order diminishes liquidity by taking depth from the order book.

Similarly, the risks associated with limit orders and market orders complement one another. A limit order avoids price risk, since the order executes at the limit price or better. In exchange, it carries execution risk, since it is not known when, or even if, it will execute. In fact, a limit order generally fails to execute precisely when a trader would, *ex post*, have most liked it to execute, and it executes when a trader would, *ex post*, have least liked it to execute.

To illustrate, consider a trader placing a limit order to buy 500 shares of XYZ at the best bid price of \$50.00. If good news about XYZ is released before the limit order is filled, the price rises and the limit order will not be filled, so the trader will have missed the chance to buy XYZ before the good news. Conversely, assume that after placing the limit order, bad news about XYZ is released. The price of XYZ drops and the order will be filled (assuming the trader did not cancel the order in time). Why, then, does a trader place a limit order? The advantage of a limit order is

that, conditional on execution, it is filled at a lower price (for buys) than if the trader had placed a market order (and at higher prices for sales).

A trader who needs to execute immediately submits a market order or marketable limit order. The cost is paying a higher price (or receiving a lower price), relative to a limit order trader, in the form of a bid–ask spread and/or price impact. A market order carries price risk, since it is not guaranteed a set price. However, it avoids execution risk, since there is no uncertainty related to execution timing or execution failure.

For markets without organized exchanges, like over-the-counter (OTC) or negotiated markets, analogs to bid and ask prices and limit order books exist. In an OTC market different dealers may quote different prices for a given asset. An agent wishing to trade an asset will generally contact a number of dealers to obtain quotes. The lowest offer to sell and the highest bid to buy correspond to the observed best bid and offer. There may be undisplayed liquidity in the market if other dealers exist that have not been contacted by the agent. Thus, it is generally impossible to see the entire implicit limit order book in OTC markets.

A useful concept is the *shadow price*, p_t^*, which gives the fair-market valuation of the security, absent news about future trades. The shadow price is not directly observable and, in some cases, we use the midpoint price as a proxy. A *perfectly liquid* market is one in which the transaction price equals the shadow price at every time point.

Simulated or back-tested portfolio risk and performance measurement depend crucially on historical records of transaction prices. Implicit in many studies is the assumption that the portfolio manager can implement trades at the historically observed prices. Consider an analysis of the risk and expected return of a portfolio strategy. Typically, an artificial set of *paper portfolio returns* is generated by a historical simulation based on the record of transaction prices.

A common finding is that paper portfolios generated by historical simulation outperform actual portfolios, having both higher average returns and lower risk. This is true even in situations where the paper and actual portfolios are run simultaneously and therefore are not subject to in-sample overfitting. This difference in performance is called *implementation shortfall* (Perold 1988). Implementation shortfall is due to the fact that the paper portfolios typically assume zero transaction costs, no price impact, and infinite liquidity at the observed transaction prices.[2] In particular, the implementation shortfall is a measure of the disparity between the shadow price and the transaction price.

[2] Some exceptions that attempt to incorporate bid–ask spreads or price impact include Schultz (1983), Stoll and Whaley (1983), Ball et al. (1995), Knez and Ready (1996), Korajczyk and Sadka (2004), and Chen et al. (2005).

In addition to the transaction price, p_t, the transaction type (buy or sell) and volume play essential roles in the measurement of transaction costs and liquidity risk. An important quantity is *order imbalance*, which is the signed volume of a trade, $\text{OI}_t = D_t \times V_t$, where V_t is the volume associated with trade t and D_t is an indicator variable equal to $+1$ for a buyer-initiated trade and -1 for a seller-initiated trade.

12.2 Measuring Transactions Cost

We consider two important facets of the cost of trading. The first is the *bid–ask spread*, which is the cost of an instantaneous round-trip purchase. Models of the bid–ask spread are generally based on a statistical analysis of transaction and quote data. The second is *price impact*, which is the effect of trading on price. We review several modeling approaches that depend on asymmetric information and adverse selection.

For both the bid–ask spread and price impact we discuss several measures that can be used when high-frequency data are available. We also consider measures that can be used in the more typical situation when only low-frequency data can be obtained.

The relationship between the bid–ask spread and market impact is complex and not completely understood. We provide some insight below in section 12.3.1, in connection with the analysis of the commonality of liquidity shocks to pools of assets.

12.2.1 Measuring the Bid–Ask Spread

12.2.1.1 Spread Measures Using High-Frequency Data

Quoted bid–ask spreads vary across stocks and over time. Figure 12.3 shows time series of the monthly average quoted spreads for stocks traded on the New York Stock Exchange (NYSE) over the period January 1983 to December 2000. For each month t, the quoted spread for firm i is defined by:

$$\text{Qspread}_{it} = \frac{1}{n_{it}} \sum_{j=1}^{n_{it}} s_{jit}^{\text{prop}},$$

where s_{jit}^{prop} is the proportional spread at the time of the jth trade of asset i in month t and n_{it} is the number of trades of asset i in month t.

For each month t, the market average (across firms) quoted spread is given by

$$\text{Qspread}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \text{Qspread}_{it},$$

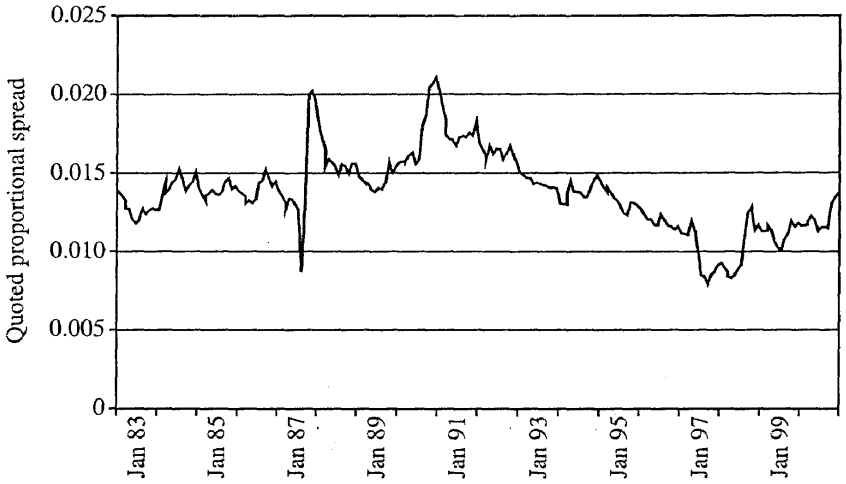

Figure 12.3. The time-series properties of the average proportional quoted spread, on a monthly basis, for stocks traded on the NYSE over the period January 1983–December 2000.

where n_t is the number of firms for which we have observations of Qspread_{it} in month t.

Over the period studied, the cross-sectional average quoted spread ranges between 0.7% and 2.1%, with a time-series mean of 1.4%. There are noticeable changes in quoted spread, particularly around the 1987 stock market crash: the months with the highest proportional quoted spreads are October and November 1987. There is a large increase in late 1990, and again in August and September of 1998 during the Russian ruble and Long Term Capital Management crises. There is a large decline in June and July of 1997 that coincides with the change in the minimum price increment for NYSE listed stocks from one-eighth of a dollar to one-sixteenth of a dollar. Between these periods of dramatic movement, there is a reasonable amount of persistence in the average quoted proportional spread.

While trades in many markets are executed at the quoted bid or ask prices, other markets have "hidden liquidity." This may be due to the fact that small limit orders are not displayed or because floor brokers may compete with market makers. In such instances, we may observe *price improvement* (see Petersen and Fialkowski 1994), which occurs when trades take place inside the quoted bid and ask prices. A measure of spread that takes price improvement into account is the *effective half spread*, defined as the absolute value of the difference between the transaction price and the midpoint price:

$$\text{es}_t = |p_t - \text{mid}_t|.$$

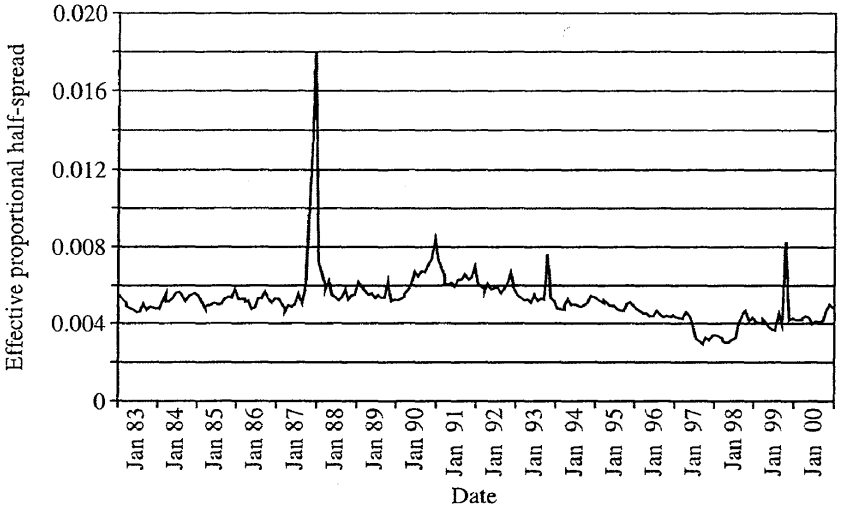

Figure 12.4. The time-series properties of the average proportional effective half spread.

The *proportional effective half spread* is defined as the *effective half spread* divided by the midpoint price:

$$\mathrm{es}_t^{\mathrm{prop}} = \frac{\mathrm{es}_t}{\mathrm{mid}_t}.$$

The effective half spread should be doubled before it is compared with a quoted spread. Figure 12.4 shows the time series of average proportional effective half spreads for NYSE firms.

For each month t, the effective half spread for firm i is the average effective spread associated with each trade of the asset within that month:

$$\mathrm{Espread}_{it} = \frac{1}{n_{it}} \sum_{j=1}^{n_{it}} \mathrm{es}_{jit}^{\mathrm{prop}},$$

where $\mathrm{es}_{jit}^{\mathrm{prop}}$ is the proportional effective spread for the jth trade of asset i in month t.

As above, we calculate the market average effective half spread by averaging cross-sectionally:

$$\mathrm{Espread}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathrm{Espread}_{it},$$

where n_t is the number of firms for which we have observations of $\mathrm{Espread}_{it}$ in month t.

Over the period studied, the market average proportional effective half spread varied between 0.3% and 1.8%, with a time-series mean of

0.5%. The average effective half spread is less than half of the average quoted spread, which is consistent with price improvement occurring on the NYSE. As for the quoted spread, there is a very large spike in the effective spread around the 1987 crash as well as an increase in late 1990 and in late 1998 and 1999. There is also a drop in June and July of 1997.

Schultz (2001) estimates effective spreads in the corporate bond market for a sample of 61,328 secondary market trades in investment-grade corporate bonds over the period January 1995–April 1997. Schultz regresses the difference between the trade price and an estimate of the prevailing bid price on an indicator variable, D_t, equal to +1 for a buyer-initiated trade and −1 for a seller-initiated trade. The regression coefficient yields an estimate of the effective spread. For this sample of bonds the estimated effective spread is 0.3%, which does not seem to be related to the credit rating of the bonds. The effective spreads are declining in trade size and are smaller for traders that are more active in the market. Note that an effective spread can be compared directly with a quoted spread, and it should be halved before it is compared with an effective half-spread.

12.2.1.2 Spread Measures Using Low-Frequency Data

The measures of liquidity discussed above rely on the availability of intraday prices and quotes. The estimators are data intensive and can be carried out only over the recent time period for which tick-by-tick intraday data are available.

There are a number of liquidity measures that are based on daily data. For example, Qspread and Espread can be estimated using daily closing bid-ask spreads. Liquidity measures that rely on low-frequency data may not be surrogates for trading cost estimates used by traders with access to intraday data. However, these measures are useful to those interested in back-testing strategies over periods or in markets for which intraday data are not available. In this section we discuss some additional measures that can be estimated with daily data.

Roll (1984) suggests a spread estimator that can be applied in situations for which direct spread data are not available, but for which transaction prices are available. Roll shows that when the chances of transacting at the bid and ask prices are each 50%, the covariance of successive price changes is determined by the spread, s,

$$\text{cov}(p_t - p_{t-1}, p_{t-1} - p_{t-2}) = -\tfrac{1}{4}s^2,$$

so that the spread can be estimated from the first-order autocovariance of changes in price:

$$s = 2[-\operatorname{cov}(p_t - p_{t-1}, p_{t-1} - p_{t-2})]^{1/2}. \tag{12.2}$$

Even though the population autocovariance should be negative in the Roll model, sample autocovariances may turn out to be positive. A common solution in this case is to estimate the spread using

$$\hat{s} = \begin{cases} 2[-\widehat{\operatorname{cov}}(p_t - p_{t-1}, p_{t-1} - p_{t-2})]^{1/2} & \text{if } \widehat{\operatorname{cov}} \leqslant 0, \\ 0 & \text{if } \widehat{\operatorname{cov}} > 0. \end{cases}$$

The Roll spread estimator can be applied to intraday data but it has typically been applied to daily data.

Bao et al. (2008) estimate the negative first-order autocovariance, $-\widehat{\operatorname{cov}}(p_t - p_{t-1}, p_{t-1} - p_{t-2})$, for a sample of 1,249 corporate bonds from April 2003 to December 2007, using both transaction data and daily data. They compare $\hat{s}$ with the actual bid-ask spread on the bonds and find that $\hat{s}$ is substantially larger. Thus, the autocovariance-based spread measure seems to be estimating not only the spread but the price impact induced by trades that are larger than the depth at the BBO. Bao et al. (2008) find that the covariance is smaller in absolute value for larger trades, and larger in absolute value when prices are declining.

Hasbrouck (2004, 2009) derives a Bayesian estimator of the Roll model that imposes the prior that the spread is positive. This addresses the problem caused by positive sample estimates of the autocovariance of price changes. The model is estimated using a Gibbs sampler. Hasbrouck evaluates the original Roll estimator and the Gibbs estimator by comparing their correlations with estimates of Espread using intraday data. The study runs from 1993 to 2005 and covers 300 stocks, half taken from the NYSE and AMEX and half taken from the NASDAQ. Hasbrouck finds that the correlation between the Roll estimator and Espread is 0.88, while the correlation between the Gibbs estimator and Espread is 0.97.[3] Thus, the Gibbs sampler estimate seems to be a better proxy than the moment-based $\hat{s}$.[4]

Holden (2009) extends the Roll model to accommodate data on days for which there is no trading. (Certain data vendors, such as the Center for Research in Security Prices, use the closing bid-ask midpoint.) He investigates versions of the Roll estimator that substitute alternative

[3] The Roll, Gibbs, and Espreads measures are estimated for each security over periods of one year. The correlation estimates are from a panel: they are based on a data set whose observations are indexed by company and year.

[4] Estimates are available from Joel Hasbrouck's Web site at http://pages.stern.nyu.edu/~jhasbrou/.

estimates of the spread when the serial covariance of price change is positive. Holden also derives a set of spread measures based on the frequency of closing prices at alternative price increments. These "effective tick" measures have high correlations with Espread.

Lesmond et al. (1999), hereafter referred to as LOT, develop a measure of effective spreads based on the insight that illiquidity impedes trading. It is often the case that on days when no trade has occurred, the reported closing price is the previous day's closing price. Thus, the frequency of days for which the closing price change is zero is a proxy for the number of days without trading. The measure of total cost (effective spread plus commission) in LOT is derived from a limited dependent variable estimator. They show that the LOT estimator is highly correlated with, but smaller than, the sum of the quoted spread and commissions.

Chen et al. (2007) study the relation between liquidity and credit spreads in the corporate bond market. Their sample includes over 4,000 noncallable corporate bonds over the period 1995–2003. They estimate three alternative measures of liquidity: the LOT estimator, an estimator based on the frequency of stock price changes equal to zero, and the proportional bid–ask spread of the bond. They find that credit spreads are significantly correlated with all the liquidity measures for both investment-grade and speculative-grade bonds, except for the "zeros" measure for speculative-grade bonds.

Goyenko et al. (2009) compare the performance of a number of low-frequency spread measures by studying their cross-sectional and time-series correlations with liquidity measures estimated using high-frequency data. The measures studied include the Roll estimator and its Bayesian extension by Hasbrouck (2004, 2009), the Holden estimator, and variants of the effective tick measure and the LOT estimator. They find that all of the low-frequency spread measures are highly correlated with the high-frequency measures. The low-frequency measures that have the highest correlation with the high-frequency measures are the Holden estimator, the effective tick model, and a variant of the LOT measure.

12.2.2 Measuring Price Impact

12.2.2.1 Price-Impact Measures Using High-Frequency Data

Price impact can have either a temporary or a permanent effect on transaction prices. The distinction depends on whether the trade changes the market's assessment of the underlying value of the asset or merely causes a temporary price movement. In the former case, the trade reveals

value-relevant information held by the initiator of the trade. We now consider the effects of information asymmetry on the costs of trading.

We begin with the model of Glosten and Milgrom (1985), in which there are two types of traders: liquidity traders and informed traders. Only market orders are allowed and there is one trade per period. Every trade is of unit size and is mediated by the market maker. The shadow price, p_t^*, is the valuation by the market maker before the time-t trade is observed. An informed trader knows that it is misvalued, say by ω. The market maker cannot distinguish between orders submitted by an informed trader and those of a liquidity trader, whose trades contain no information about the true value of the security. The market maker knows only the proportion, $\Pr(I)$, of informed trades and the potential mispricing, ω. Note that $\Pr(I)$ is also the probability that any given trade is an informed trade.

For simplicity, and in order to highlight the relationship between information and liquidity, Glosten and Milgrom assume that the market maker's inventory and opportunity costs are zero. It follows that the only role of the bid–ask spread is to compensate the market maker for the adverse-selection effect that arises from the presence of informed traders. The market maker sets his bid–ask prices, p_t^{b} and p_t^{a}, so that, contingent on a buy or sell order, the quoted price equals the expected value of the security, given the probability of an informed trade:

$$p_t^{\mathrm{a}} = p_t^* + E[p_t - p_t^* \mid \text{buy}] = p_t^* + \Pr(I)\omega,$$
$$p_t^{\mathrm{b}} = p_t^* + E[p_t - p_t^* \mid \text{sell}] = p_t^* - \Pr(I)\omega.$$

The Glosten–Milgrom model has the property that in the absence of external information shocks, the market maker's new shadow price, p_{t+1}^*, equals the previous transaction price, p_t^{a} or p_t^{b}. This has two interesting consequences. First, since the new bid and ask prices are bracketed around the new shadow price, transaction price changes have zero autocovariance: there is no bid–ask bounce in this model. Second, the model makes clear that the Roll (1984) autocovariance estimate measures only a component of the bid–ask spread: the portion due to market making costs and dealer monopoly rents. The component due purely to adverse selection (the component captured by the Glosten–Milgrom model) does not induce negative autocorrelation.

Both the Roll and Glosten–Milgrom models contain fundamental insights about the components of transaction costs. However, the Roll model does not take account of the presence of informed traders and the Glosten–Milgrom model constrains informed traders to submit a single, unit-size buy or sell order. Kyle (1985) develops an alternative model of

the adverse-selection component of the transaction costs, allowing for varying order size. In the Kyle model, as in the Glosten–Milgrom model, a competitive market maker sets a price schedule as a function of order imbalance, subject to a zero-profit condition and to the market maker's inability to distinguish between informed and liquidity traders. There is one informed trader and a collection of uninformed liquidity traders who submit orders simultaneously. The market maker accommodates the net order imbalance, which is the difference between total buy orders and total sell orders. In the Kyle model the market maker's bid–ask prices take the form of a price schedule, with higher ask prices and lower bid prices for larger absolute order imbalances. An informed trader chooses the optimal order size based on the market maker's price schedule and the quality of the information signal. The equilibrium in this model is a price schedule in which the change in transaction price is proportional to order imbalance OI:

$$p_t - p_{t-1} = \lambda \times \mathrm{OI}_t \,. \tag{12.3}$$

The price-impact coefficient, λ, is increasing in the precision of the informed trader's information and decreasing in the volatility of the noise trader's order flow.

A weakness of some of the models we have considered is the lack of a relationship between order flow and bid and ask prices. Consider, for example, the Roll model, in which the market maker sets a fixed bid–ask spread to compensate for the risk of holding inventory of the security. Suppose that, by chance, over a particular time period, there is a large excess of sell orders over buy orders. Then, over that period, the market maker's inventory will grow. It seems natural that the market maker's risk compensation should also grow with the size of his inventory. Madhavan and Smidt (1993) develop a model with a dynamic correction for the effect of order flow on bid and ask prices. In addition to the Kyle-type information-based spread, the market maker adjusts the bid–ask spread dynamically in order to control his inventory position.

Trading a block larger than the depth at the inside quotes moves the price adversely, pushing prices up with a purchase and down with a sale; recall that this movement in price is called the *price impact.* In the Kyle (1985) model, the only cost that the competitive market marker faces arises from the adverse selection inherent in trading against an informed trader. In reality the market maker needs to recover other costs such as inventory carrying costs, back-office costs, labor, and clearing fees. Additionally, if the market for market-maker services is not completely competitive, the market maker may earn monopoly profits. Glosten and Harris (1988) develop a model of the bid–ask spread that incorporates

both the adverse-selection problem faced by a market maker trading with informed traders and the non-information-based costs of market making. The adverse-selection problem leads to permanent price impacts as in Kyle (1985). In contrast, market-maker transaction costs lead to transitory price impacts that generate negative serial correlation in transaction prices.

Glosten and Harris also develop an econometric model to estimate the adverse-selection and market-maker carrying-cost components. Their econometric technique estimates the trade direction, D_t, for each trade and incorporates discreteness of prices induced by a minimum price increment, or *tick size*. Variants of this model that rely on a separate trade classification algorithm (so that trade direction estimation is not part of the model) are commonplace.

Sadka (2006) develops an extension of the specification in Glosten and Harris. Consider the market maker's expected value of a security, mv_t, conditional on the information set available at the time of a trade, t:

$$\text{mv}_t = E_t[\text{mv}_{t+1} \mid D_t, V_t, y_t], \tag{12.4}$$

where V_t is the volume associated with trade t, D_t is an indicator variable equal to $+1$ for a buyer-initiated trade and -1 for a seller-initiated trade, and y_t is public, non-trade-related information.[5] To determine D_t, Sadka classifies a trade whose price is above the midpoint of the quoted bid and ask prices as being buyer initiated and a trade whose price is below the midpoint as being seller initiated. A trade whose price is at the midpoint is discarded from the estimation.

Glosten and Harris (1988) and Sadka (2006) assume that price impact has a linear functional form. Huberman and Stanzl (2004) show that the permanent component of the price-impact function must be linear in order to rule out quasi-arbitrage opportunities.[6] Sadka (2006) posits four components of price impact: permanent and transitory sensitivities to trade type (buy and sell), denoted by Ψ and $\bar{\Psi}$, and permanent and transitory sensitivities to order flow, denoted by λ and $\bar{\lambda}$.

To estimate the permanent price effects, Sadka follows the formulation proposed by Glosten and Harris and assumes that mv_t takes the linear form

$$\text{mv}_t = \text{mv}_{t-1} + D_t[\Psi + \lambda V_t] + y_t, \tag{12.5}$$

---

[5] In this case, the market maker's time-t price, mv_t, takes account of a trade that has just occurred at time t.

[6] Since traders do not know their execution prices with certainty, a pure arbitrage opportunity is not feasible. Instead, they search for quasi-arbitrage opportunities, which are unbounded price manipulations for which the limit of the Sharpe ratio is infinite.

where Ψ and λ are the trade-type and order-flow permanent price-impact costs, respectively. Equation (12.5) describes the innovation in the conditional expectation of the security value through new information that is trade related (D_t, V_t) or not trade related (y_t). Notice that new information has a permanent impact on expected value.

The (observed) transaction price, p_t, can be written as

$$p_t = \mathrm{mv}_t + D_t[\bar{\Psi} + \bar{\lambda} V_t], \tag{12.6}$$

where $\bar{\Psi}$ and $\bar{\lambda}$ are transitory effects, since they affect p_t but not p_{t+1}. Taking first differences of p_t (equation (12.6)) and substituting $\mathrm{mv}_t - \mathrm{mv}_{t-1}$ from equation (12.5) we have

$$p_t - p_{t-1} = \Psi D_t + \lambda D_t V_t + \bar{\Psi}(D_t - D_{t-1}) + \bar{\lambda}(D_t V_t - D_{t-1} V_{t-1}) + y_t, \tag{12.7}$$

where y_t is the unobservable residual due to non-trade-related information.

Equation (12.7) allows us to interpret the model parameters Ψ and λ as the sensitivities of price to trade type and order flow. Similarly, the model parameters $\bar{\Psi}$ and $\bar{\lambda}$ are the sensitivities of price to *change* in trade type and *change* in order flow. Equation (12.7) assumes that the market maker revises expectations according to the total order flow observed at time t. However, there is documented predictability in order flow (Hasbrouck 1991a,b; Foster and Viswanathan 1993). For example, breaking large trades into smaller trades to reduce price impact creates autocorrelation in order flow. The value-relevant equation (12.7) is adjusted to account for the predictability in order flow. In particular, the market maker is assumed to revise the conditional expectation of the security value according to only the *unanticipated* order flow rather than the entire order flow at time t. The unanticipated order flow, denoted by $\varepsilon_{\lambda,t}$, is calculated as the fitted error term from a five-lag autoregression of order flow, $D_t \times V_t$ (after computing $\varepsilon_{\lambda,t}$, the unanticipated sign of the order flow, $\varepsilon_{\Psi,t}$, is calculated while imposing normality of the error term, $\varepsilon_{\lambda,t}$ (see Sadka (2006) for more details)). Therefore, equation (12.7) translates to

$$p_t - p_{t-1} = \Psi \varepsilon_{\Psi,t} + \lambda \varepsilon_{\lambda,t} + \bar{\Psi}(D_t - D_{t-1}) + \bar{\lambda}(D_t V_t - D_{t-1} V_{t-1}) + y_t. \tag{12.8}$$

Somewhat counterintuitively, the empirical literature documents an inverse relationship between the permanent order-flow sensitivity, λ, and the size of the order block. This is probably due to the fact that information about the block reaches the market in advance of the actual trade (Nelling 2003). Therefore, the block trade appears to have a small price impact when price change is measured relative to the previous trade. In

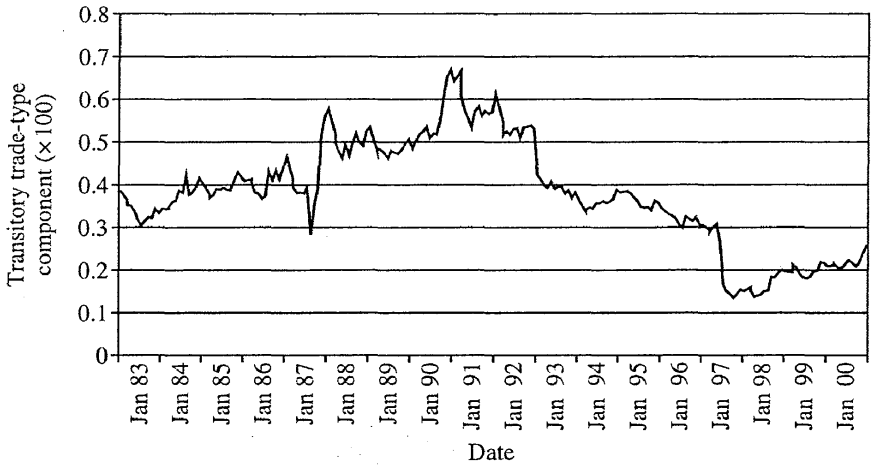

Figure 12.5. A plot of the monthly average of the transitory trade-type component of the bid–ask spread expressed as a percentage of the beginning-of-month price.

light of this, Sadka segregates block trades (trades above 10,000 shares) in the estimation. The model in equation (12.8) is estimated separately for each stock every month using ordinary least squares (including an intercept). In Glosten and Harris the primary parameters of the model are the transitory trade-type sensitivity and the permanent order-flow sensitivity. These are generally interpreted as the components of price change due to market-making costs and adverse selection. We focus on these two components of price impact here.

Figure 12.5 is a plot of the monthly average of the transitory component expressed as a percentage of the beginning-of-month price. From equation (12.6), this component is essentially half of the effective bid-ask spread. Because of this we would expect $\bar{Y}/p$ to behave much like the proportional spread measures discussed above. Indeed, the transitory sensitivity behaves very much like the quoted spread measure, except that it is, on average, smaller. This is to be expected since (a) $\bar{Y}/p$ is a half-spread measure while Qspread measures the full spread and (b) $\bar{Y}/p$ measures an effective spread while Qspread is a quoted spread. The transitory component, $\bar{Y}/p$, behaves like the effective spread measure, Espread, with the exception that Espread has a more pronounced peak during the crash of 1987.

Figure 12.6 is a plot of the monthly average of the permanent component expressed as a percentage of the beginning of month price. The quantity λ/p shows large jumps around the 1987 crash and the 1998 Long Term Capital Management and Russian ruble crisis. There are also

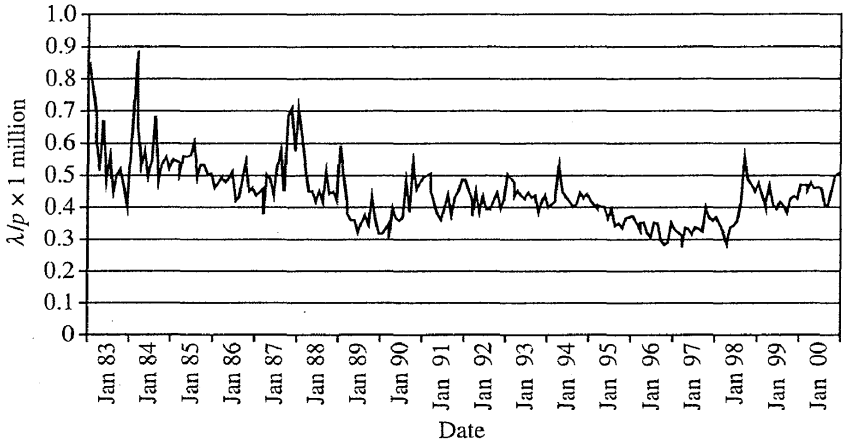

Figure 12.6. A plot of the monthly average of the permanent order-flow component of the bid-ask spread expressed as a percentage of the beginning-of-month price.

some large spikes in February 1983 and February 1984, which are more difficult to explain.

The Kyle (1985) and Glosten and Harris price-impact measures require transaction prices and a method of classifying trades as buyer initiated or seller initiated. Typically, trade classification algorithms use quote and transaction price data and do not require depth data. In markets for which it is possible to observe the limit order book, we can measure the price impact for a hypothetical trade by calculating the average price paid to fill the order by placing one market order and sweeping through the limit order book. Farmer et al. (2004) and Burghardt et al. (2006) calculate such "sweep-to-fill" price-impact measures for individual equities and E-mini S&P futures contracts, respectively.

12.2.2.2 Price-Impact Measures Using Low-Frequency Data

Amihud (2002) considers the average daily ratio of the absolute value of stock return to dollar volume. Amihud argues that this measure "can be interpreted as the daily price response associated with one dollar of trading volume, thus serving as a rough measure of price impact." In a Kyle-type model, the average ratio of price change to order imbalance converges to the price-impact coefficient. Using the absolute return in the Amihud measure replaces the numerator and denominator with quantities that are upward biased. If the price change is due to information revealed by the order imbalance and to other news, then price changes

are given by the Kyle model (12.3) plus non-trade-related news:

$$p_t - p_{t-1} = \lambda \times \mathrm{OI}_t + \varepsilon_t, \tag{12.9}$$

where ε_t represents the price reaction to non-order-related news. Thus, assuming that $E(\varepsilon_t)$ and $E(\mathrm{OI}_t)$ equal zero, the average ratio of the absolute price change to the absolute order imbalance converges to

$$\sqrt{\lambda^2 + \frac{\sigma_\epsilon^2}{\sigma_{\mathrm{OI}}^2}},$$

which is larger than λ unless the variance of ε is zero. However, the denominator in the Amihud measure is not the absolute value of OI but is instead the dollar volume. Since OI equals buyer-initiated volume minus seller-initiated volume while volume is the sum of the two, the Amihud measure has a denominator that is larger than the absolute value of OI. Since both the numerator and the denominator are upward biased, the net effect is indeterminate.

A positive feature of the Amihud measure is that it can be calculated using low-frequency data, whereas any measure using an estimate of the order imbalance requires intraday data. Figure 12.7 plots the monthly estimate of the Amihud measure, averaged across firms traded on the NYSE. For each firm the measure is estimated on a monthly basis by

$$A_{it} = \sum_{j=1}^{d_t} \frac{|r_{ij}|}{\mathrm{DV}_{ij}}, \tag{12.10}$$

where r_{ij} is the return on asset i on day j of month t, DV_{ij} is the dollar volume traded in asset i on day j of month t, and d_t is the number of trading days in month t. For inclusion in the month t sample we require asset i to have observations on $|r_{ij}|/\mathrm{DV}_i$ for at least fifteen days. The figure plots the cross-sectional average measure

$$\bar{A}_t = \sum_{i=1}^{n_t} \frac{A_{it}}{n_t},$$

where n_t is the number of firms for which data are available in month t. The Amihud measure has local peaks at many of the same times as the previous measures. However, the October 1987 crash stands out less prominently than it does for the other measures. In addition, the Amihud measure seems to have more month-to-month variability than the previous measures. Hasbrouck (2005) finds that A_{it} exhibits a large amount of kurtosis, so the high variability might be due to large outliers.

Downing et al. (2008) study the pricing of illiquidity in the corporate bond market using the Amihud measure and a variant that uses the

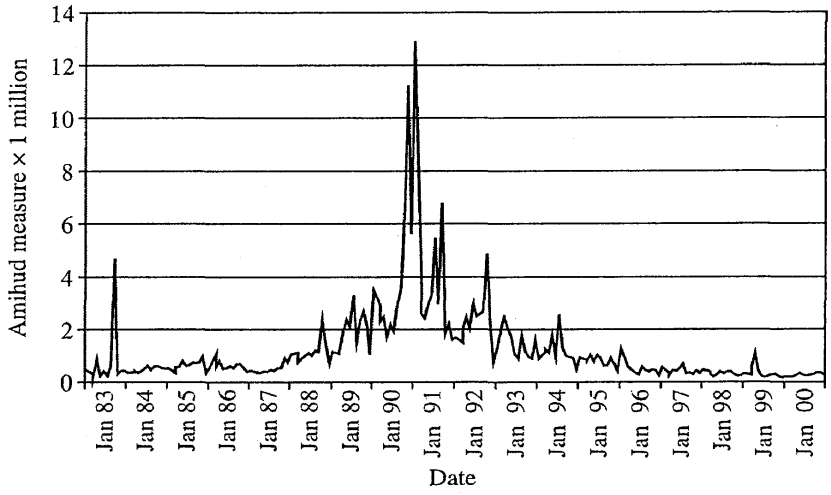

Figure 12.7. The monthly estimate of the Amihud liquidity measure, averaged across firms traded on the NYSE.

spread between the high and low prices over the observation interval instead of the period return in the numerator of (12.9). They find that the absolute level of bond liquidity as well as the covariance of bond returns with aggregate liquidity command a return premium in the corporate bond market.

Pástor and Stambaugh (2003) measure liquidity as the volume-related daily return reversal, γ, in the regression

$$r^a_{ij+1t} = \theta_{it} + \varphi_{it} r_{ijt} + \gamma_{it} \operatorname{sign}(r^a_{ijt}) \, \mathrm{DV}_{ijt} + \varepsilon_{ij+1t},$$

where $r^a_{ijt} = r_{ijt} - r_{mjt}$, r_{ijt} and r_{mjt} are the returns on asset i and the market portfolio, m, respectively, on day j of month t, and DV_{ijt} is the dollar volume traded in asset i on day j of month t. The coefficient γ is expected to be negative since large volume on day j will lead to temporary price movements that will reverse themselves on day $j + 1$. Pástor and Stambaugh (2003) scale the average values of the volume-related return reversals,

$$\hat{\gamma}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \hat{\gamma}_{it},$$

by the market capitalization of firms in month t relative to the market capitalization of firms in August 1962 (the beginning of their sample period). This yields an estimate

$$\hat{\gamma}^*_t = \hat{\gamma}_t \times \frac{\mathrm{mc}_t}{\mathrm{mc}_0},$$

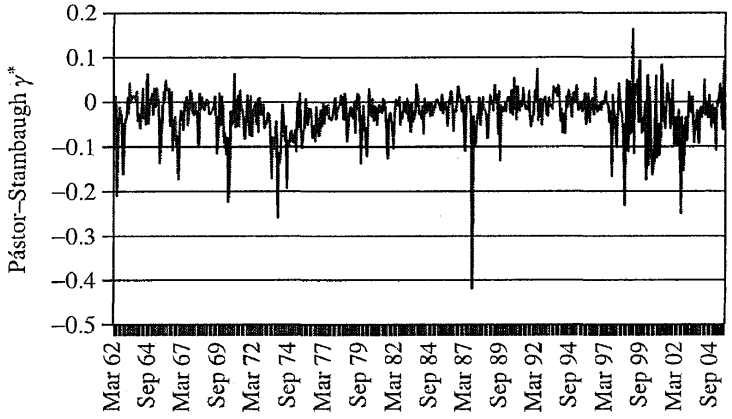

Figure 12.8. Pástor-Stambaugh γ^*.

where mc_t is the aggregate market capitalization of firms in their sample in month t. Their aggregate series, plotted in figure 12.8, shows pronounced declines in liquidity around the 1987 crash, the Long Term Capital Management and Russian ruble crises, and the 1973 oil embargo.

Goyenko et al. (2009) compare the performance of a number of low-frequency price-impact measures by studying their cross-sectional and time-series correlations with liquidity measures estimated using high-frequency data. The measures studied include A_{it}, the Amihud (2002) measure, γ from Pástor and Stambaugh (2003), and the Amivest liquidity ratio, which is constructed from $\text{DV}_{ij}/|r_{ij}|$, the inverse of the variable used in A_{it}. They find that A_{it} has significant correlation with the high-frequency price-impact measures. The Amivest measure and γ are not highly correlated with the high-frequency price-impact measures.

12.2.3 Other Variables Correlated with Liquidity

A number of variables are correlated with the level of an asset's liquidity. It is natural to expect liquid assets to have high turnover (volume divided by the number of units outstanding). For each stock i, turnover for month t is defined as

$$\text{Turnover}_{it} = \sum_{j=1}^{d_t} \frac{V_{ij}}{\text{SO}_{it}}, \tag{12.11}$$

where Turnover_{it} is the turnover in asset i for month t, V_{ij} is the trading volume in asset i for day j, d_t is the number of trading days in month t, and SO_{it} is the number of shares outstanding for asset i in month t. The cross-sectional average of Turnover_{it} gives us a measure of market-wide turnover. Figure 12.9 plots average monthly turnover for NYSE stocks.

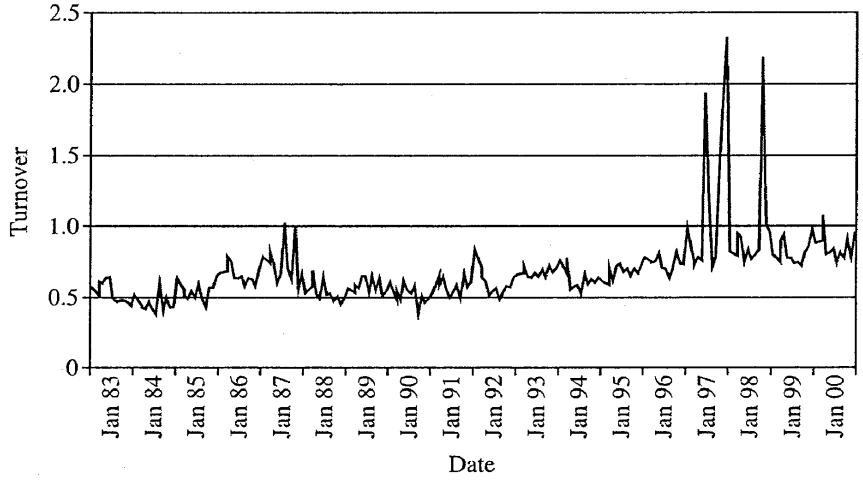

Figure 12.9. Plot of the average monthly turnover for NYSE stocks, defined as volume divided by shares outstanding.

While cross-sectional differences in average turnover are likely to be correlated with cross-sectional differences in liquidity, the time-series plot tends to show peaks in turnover when other metrics of liquidity are low. This occurred, for example, during the 1987 crash.

12.3 Statistical Properties of Liquidity

Many portfolio managers are subject to net investments that may be negatively correlated with both the portfolio return and the liquidity of the assets. In other words, they may experience redemptions when the assets in the portfolio are least liquid. Their portfolio positions and trading strategies should take into account the expected liquidity of assets as well as the risk associated with changes in assets' liquidity. The cost of liquidity in real portfolio trades is particularly large during market downturns, and this can be missed by risk estimates based on paper portfolio returns.

From a portfolio perspective, it is important to determine whether liquidity poses a systematic risk. An idiosyncratic shock to the liquidity of a single asset is less risky than a shock that affects a pool of assets. Additionally, it is important to determine the persistence of a liquidity shock. A shock that is transitory (relative to the flow of assets into and out of the portfolio) is of less concern than one that exhibits strong persistence.

12.3.1 Commonality of Liquidity Shocks

There are a number of papers that investigate whether liquidity shocks are common across assets. Chordia et al. (2000) find strong evidence of commonality across assets in liquidity, measured by quoted spreads, effective spreads, and depth. Their sample includes intraday transaction data for ordinary shares of NYSE firms over the year 1992. They require that firms be continuously listed over the year and that they have trading on at least ten days. They exclude firms who split their shares or have a stock dividend. There are 1,169 firms in the sample. They measure the common component by regressing daily changes in asset liquidity on changes in "market" liquidity, defined as the average liquidity across assets. While statistically significant, market-wide liquidity explains only a small fraction of the variability in liquidity across firms.

Eckbo and Norli (2002) take a similar approach at a monthly horizon, using turnover, spread, and price impact as measures of liquidity. They study NYSE, AMEX, and NASDAQ stocks over the period 1963–2000. Like Chordia et al., they find significant commonality in liquidity across assets.

Hasbrouck and Seppi (2001) study commonality in liquidity across the thirty stocks in the Dow Jones Industrial Average for the year 1994. Using high-frequency intraday data they find that liquidity measures, such as spreads, depth, and the "slope" of the supply curve, each have common systematic factors.

These studies and others provide clear evidence that most measures of liquidity have common components across assets. In addition there may be commonality across liquidity measures. This may be due to the fact that the measures are estimates of similar underlying quantities. For example, Qspread, Espread, and $\bar{\Psi}$ are all spread measures. Alternatively, the measures might estimate different aspects of liquidity that should, in theory, be correlated. For example, market-making costs reflected in $\bar{\Psi}$ should have an effect on the pool of informed traders, whose information precision is reflected in the permanent price-impact coefficient λ (see Glosten 1987). Korajczyk and Sadka (2008) estimate factor models for monthly observations on eight different measures of liquidity: the high-frequency measures of bid-ask spread, Qspread_{it} and Espread_{it}; the permanent and transitory components of price impact defined in Sadka (2006), Ψ_{it}, λ_{it}, $\bar{\Psi}_{it}$, and $\bar{\lambda}_{it}$; the lower-frequency Amihud (2002) ratio of absolute return to volume, A_{it}; along with Turnover_{it}. In addition to the factor models for each liquidity measure, they estimate a factor model pooled across the cross-sectional sample of stocks and liquidity measures. The sample consists of 4,055 NYSE traded stocks over the

period January 1983–December 2000. For the eight liquidity measures, a one-factor model explains between 4% and 25% of the liquidity of individual assets, on average. A three-factor model explains between 12% and 55% of the liquidity of individual assets, on average. They find significant correlations between the factors extracted from the individual liquidity measures.

Liquidity is correlated with asset returns, as shown, for example, in Chordia et al. (2001) and Korajczyk and Sadka (2008). This implies that liquidity tends to dry up when asset returns are negative. This may be precisely when some portfolio strategies are forced to liquidate assets, thus exacerbating the liquidity problem.

12.3.2 Persistence of Liquidity Shocks

Monthly estimates of various liquidity measures generally show some time-series persistence, punctuated by occasional large jumps. We estimate first- and second-order autoregressive (AR(1) and AR(2)) models for several of the liquidity measures we study. The first-order autocorrelation of the measures ranges between 0.6 (for turnover) and 0.98 (for $\bar{\Psi}$) with the exception of the scaled Pástor–Stambaugh estimate $\hat{\gamma}_t^*$, whose first-order autocorrelation is 0.19. The same type of autocorrelation structure is evident in the factors extracted from the cross section of liquidity measures (see, for example, Korajczyk and Sadka 2008).

12.3.3 Jump and Event Risk

Figures 12.3–12.9 above show a high level of persistence in liquidity and also illustrate the striking fact that all of the liquidity measures exhibit evidence of large jumps. These jumps tend to occur during periods of market disruption. These disruptions include the 1987 stock market crash, the Russian financial crisis, and the period around the Gulf War in 1991. The Pástor–Stambaugh series also shows evidence of a shock to liquidity around the 1973 Arab–Israeli War.

It is clear that movements in liquidity can be correlated with asset returns. This is particularly evident during periods with downward jumps in liquidity, which are often associated with market downturns. If a portfolio is subject to redemption risk during such market disruptions, then there is additional risk induced by the correlation between portfolio flows and adverse changes in liquidity.

12.3.3.1 Headline-Generating Liquidity Crises

There is a large literature devoted to the analysis of headline-generating liquidity crises. Jorion (2000) chronicles the infamous disintegration of

Long Term Capital Management in the late summer and early autumn of 1998. This crisis, coupled with the contemporaneous ruble default, led to months of market decline and high volatility. Many believe that without the intervention of the Federal Reserve Bank of New York, the Long Term Capital Management crisis might have destroyed the world's financial systems. While spokespersons for Long Term Capital Management ascribe the fund's failure to events that were "beyond the fund's capacity to anticipate," Jorion asserts that the fund "severely underestimated its risk" and that "even if it had measured its risk correctly, the firm failed to manage its risk properly." Jorion explains how Long Term Capital Management failed to account for the dynamics of risk, described in chapter 9, using data from a lower-volatility regime to make forecasts in a higher-volatility regime. Furthermore, they ignored the asymmetric and heavy-tailed profiles of the loss distributions of their exotic, highly levered portfolio. These aspects of risk are described in chapter 10. Finally, when Long Term Capital Management encountered its first big losses in May 1998, it chose to sell off its most liquid positions because they were expected to be less profitable at the time. The firm retained only its less-liquid positions while holding inadequate capital reserves.

The causes, and even some of the effects, of the 2007–8 liquidity crisis remain obscure, despite extensive inquiry and analysis. This may be attributed to the secrecy surrounding hedge funds and to the complexity of global financial markets and securities. One of the most striking features of the crisis is the spike in interbank lending rates. On August 14, the LIBOR rate climbed to a high of over 200 basis points from its normal level of roughly 50 basis points. A contributing factor was the realization by market participants that the risk profile of exotic derivatives, such as the CDOs discussed in chapter 11, might be poorly understood. This resulted in a loss of confidence that constricted major banks and may account for at least part of the atypical risk premium associated with interbank lending rates in August 2007. Michaud and Upper (2008) decompose the risk premium[7] on interbank rates into a sum:

$$\text{rprem} = \text{credit} + \text{tprem} + \text{micro} + \text{mliq} + \text{bliq},$$

where credit, tprem, and micro are premia for the risk associated with default, term, and market microstructure, and mliq and bliq are liquidity premia. The measure mliq uses the Roll (1984) bid-ask spread estimator, which is given in equation (12.2). The term bliq is a measure of market impact obtained by regression of return onto order flow. The analysis

[7] Michaud and Upper (2008) model the interbank risk premium as the spread between LIBOR rates and rates on overnight index swaps.

in Michaud and Upper showed that both liquidity measures increased dramatically from norms of 1 or 2 basis points, during August 2007. The Roll measure mliq jumped to a high of 31 basis points, and the impact measure bliq jumped to a high of 15 basis points.

Many market-neutral hedge funds experienced substantial losses during August 2007, and quite a few went out of business. The opacity enjoyed by hedge funds precludes a careful analysis of what actually happened. However, Khandani and Lo (2007, 2008) attempt to reverse engineer the details using information from the Lipper-TASS hedge fund database and a simulation of a quantitative strategy. The authors posit that the turbulent market conditions generated margin calls that required many hedge funds to unwind their strategies simultaneously.

The events surrounding the financial market crisis beginning in 2007 were partly a credit problem and partly a liquidity problem. The housing downturn led to losses by holders of subprime mortgage obligations. However, as mentioned above, the opacity of the financial markets and the interlinkages present in them meant that it was difficult for any institution to assess the size of the risks to which any particular counterparty was exposed. This uncertainty led to the cessation of interbank lending due to the fact that it was nearly impossible to assess the credit condition of a given institution. The credit condition depended on the positions taken by that institution and on the soundness of all its counterparties, which, in turn, depended on the soundness of the counterparties' counterparties, and so on (see Gorton 2008, 2009).

The linkage between liquidity and opacity is evident in the comparison of equity and credit markets. The equity markets in 2007 and 2008 remained relatively liquid while the more opaque credit markets shut down in some instances.

12.3.3.2 Liquidity and Corporate Events

Liquidity can change around corporate events. It is often argued that stock splits increase the number of uninformed, retail investors holding the stock and, hence, increase liquidity. For some indirect measures of liquidity, there seems to be evidence to support this argument. Lamoureux and Poon (1987), Brennan and Hughes (1991), and Maloney and Mulherin (1992) document an increase in the number of shareholders, institutional ownership, the number of shares traded, dollar volume, and the number of trades following splits. Several authors find that more direct measures of liquidity decline after stock splits. Copeland (1979), Conroy et al. (1990), and Schultz (2000) find an increase in the proportional bid-ask spread after splits. Lakonishok and Lev (1987) and Gray

et al. (2003) find reduced dollar market depth and dollar trading volume subsequent to splits. Goyenko et al. (2006) find that these declines in liquidity following splits are transitory and that liquidity increases in the long run for splitting firms.

There is also evidence that spreads and depth change in the period surrounding firms' earnings announcements. Lee et al. (1993) find that spreads widen and depth decreases in anticipation of earnings releases (see also Venkatesh and Chiang 1986; Libby et al. 2002).

The timing of some of these corporate events is forecastable and the anticipated changes in liquidity can be incorporated into the risk analysis of a trading strategy. Other events are unanticipated. The changes in liquidity due to these events must be dealt with after the fact. However, most event-driven liquidity shocks seem to be relatively short-lived, thus posing less of a problem for portfolios that can postpone trading.

12.4 Optimal Trading Strategies and Transaction Costs

Trading costs can be a significant source of portfolio risk and a substantial drag on portfolio performance. Schultz (1983) and Stoll and Whaley (1983) estimate the effects of commissions and spreads on size-based trading strategies. They find that transaction costs have a large effect on the profitability of small-capitalization trading strategies, particularly those with large turnover. Ball et al. (1995) show that microstructure effects, such as bid–ask spreads, significantly reduce the profitability of a contrarian strategy. Grundy and Martin (2001) calculate that at round-trip transaction costs of 1.5%, the profits on a long–short momentum strategy become statistically insignificant. At round-trip transaction costs of 1.77%, they find that the profits on the long–short momentum strategy are driven to zero.

The importance of incorporating nonproportional price impact into the analysis of trading strategies is increasingly apparent. Knez and Ready (1996) study the price-impact effects on the profitability of a trading strategy based on the autocorrelation and cross-autocorrelation of large-firm and small-firm portfolios. They find that the trading costs swamp the abnormal returns to the strategy. Mitchell and Pulvino (2001) incorporate commissions and price-impact costs into a merger arbitrage portfolio strategy. They find that the trading costs reduce the profits of the strategy by 300 basis points per year.

Lesmond et al. (2004) and Korajczyk and Sadka (2004) study the effects of illiquidity on momentum strategies, while Chen et al. (2005) study

size, book-to-market, and momentum strategies. They find that trading costs have significant effects on the profits of the strategies they study. For example, while equal-weighted momentum trading strategies outperform value-weighted strategies before trading costs, value-weighted strategies dominate after taking account of the cost of the effective spread and price impact from a Glosten-Harris price-impact model (see Korajczyk and Sadka 2004). Korajczyk and Sadka (2004) also derive liquidity-tilted trading strategies. With a Kyle-type price-impact model and a number of simplifying assumptions, the optimal liquidity-tilted weights are proportional to value weights and are inversely proportional to the price-impact coefficient. Empirically, these liquidity-tilted portfolios provide superior performance after taking into account the cost of price impact.

A number of papers consider the problem of executing trades in a way that minimizes transaction costs. Bertsimas and Lo (1998) study the problem of minimizing the expected cost of executing an exogenously specified trade of size $\bar{S}$ over an exogenously given horizon. They obtain the best execution strategy as the solution to a dynamic optimization problem that is specified mathematically in terms of a transaction price process (that includes a random component and a price-impact term), an objective function, and constraints. The authors begin with a simple price process in which impact depends linearly on order imbalance and randomness is white noise. The change in price is given by equation (12.9). Bertsimas and Lo minimize the cost of buying $\bar{S}$ shares over the horizon $t = 1, 2, \ldots, T$ as a sequence of orders. Their objective is represented mathematically as

$$\min_{\{\mathrm{OI}_t\}} E_t\left[\sum_{t=1}^{T} \mathrm{OI}_t\, p_t\right]$$

and it is subject to the constraint that

$$\sum_{t=1}^{T} \mathrm{OI}_t = \bar{S}.$$

Using iterated applications of the Bellman equation, the authors conclude that if prices follow the simple process specified in equation (12.9), then the optimal (lowest-impact) strategy is to trade an equal number of shares at every point in time.

Bertsimas and Lo consider price processes that are more economically plausible (and more complicated) than equation (12.9). For example, they consider an extension of equation (12.9) with a noisy, temporally dependent information term that changes the rate of trading. The linear

price-impact term and price process are the same as in the equilibrium determined by the Kyle (1985) model. With the same price-impact models but with private information about the expected direction of future price changes, the optimal trading strategy can speed up or slow down trading relative to the strategy of trading equal numbers of shares each period. For example, we speed up purchases and slow down sales with a forecast of future price increases. Approaches to determining the optimal trading strategy corresponding to more general price dynamics and a nonlinear specification of market impact are derived and, in some cases, implemented. Bertsimas and Lo also consider optimal trading of multiple positions, taking into account the possibility that trades in one asset influence the prices of other assets. This is likely to be an issue for arbitrage positions in which some assets are hedges for others.

Bertsimas and Lo focus on minimizing the cost of the trade, given the exogenous constraints. Cost-minimizing strategies break orders into smaller components to reduce the effect of price impact. This exposes the trader to execution risk, primarily from two sources: (1) the equilibrium price of the asset may move adversely due to news that is unrelated to the trades being executed by the trader; and (2) the liquidity of the asset may deteriorate, making future trades more costly. The trade-off between execution costs and the first type of risk is studied by Almgren and Chriss (2000). Under assumptions very similar to those made by Bertsimas and Lo and using a price-impact function like that in Glosten and Harris (1988), Almgren and Chriss (2000) derive optimal trading strategies that explicitly trade off execution costs with the risk of adverse price movements when traders have mean-variance utility.[8] In their setting, the expected cost-minimizing strategy of Bertsimas and Lo is optimal for risk-neutral agents. Risk averse agents will liquidate the portfolios more rapidly initially, followed by slower trading. Similar results and a number of extensions are derived by Grinold and Kahn (2000), Huberman and Stanzl (2005), and Engle and Ferstenberg (2007).

In most of these analyses, the trading horizon is taken as exogenously given, but is unspecified. Huberman and Stanzl (2005) discuss the comparative statics of the determinants of the number of trades and the trading horizon. Empirically, most institutional orders are executed within a day. Breen et al. (2002) find that, in a sample of institutional orders, 92.5% are completed on the same day that trading is initiated. Thus, for most trades, the trader's horizon seems to be one trading day or less.

[8] The linear price-impact function allows for closed-form solutions, while numerical methods may be required for other functional forms.

With a one-day trading horizon, the trading strategy implied by quadratic utility, as used in Almgren and Chriss (2000), Grinold and Kahn (2000), Huberman and Stanzl (2005), and Engle and Ferstenberg (2007), implies a trading pattern different from the intraday pattern observed empirically (in Harris (1986) for example), where volume is high at the beginning and end of the trading day. It may be that the high volume at the beginning of the day is caused by traders following the optimal mean–variance trading strategy and that the high volume at the end of the day is due to some other type of trader (e.g., index funds that wish to trade at the closing price).

An alternative explanation for the observed pattern in trading volume is proposed in Hora (2006). Hora argues that mean–variance preferences over total execution costs induce a preference for early execution due, in part, to the manner in which the utility specification links risk aversion and the intertemporal elasticity of substitution (see, for example, Epstein and Zin 1989; Weil 1990). Hora specifies a cost function that depends on the implementation shortfall and its variance at each round of trading plus a term proportional to the squared unexecuted amount of the order. The optimal expected execution path is U-shaped, with high rates of execution at the beginning of trading and at the end of the trading horizon, similar to those observed empirically.

The papers discussed above analyze the appropriate trading strategy, given the order to execute a certain package of trades. Alternatively, one could consider the decision of what assets to trade, given a liquidity shock, such as a redemption by investors. Constantinides (1986) and Heaton and Lucas (1996) study portfolio decisions by a representative agent in which stocks are less liquid than bonds. The first-order effect is that investors concentrate trades in the most liquid assets, with infrequent rebalancing in the illiquid assets. Only when the agent's portfolio is sufficiently far from the optimum position (ignoring transaction costs) will the agent trade in the illiquid asset. This make sense from the standpoint of balancing trading costs with the utility loss of being far from the "optimum" position.

For institutional traders subject to asset withdrawals and financing risks, the strategy of responding to a liquidity shock by liquidating the liquid assets first has some associated risks, such as those discussed above regarding Long Term Capital Management. Selling the liquid assets first means that the remaining portfolio is less liquid. This might induce investors to "run" on the portfolio, lest they be the last investors left holding the least liquid of the assets. This problem is exacerbated if there are a number of portfolio managers holding similar positions and subject to correlated liquidity shocks. Their simultaneous trading may lead to

large price disruptions (Khandani and Lo 2007, 2008) and a significant decrease in market liquidity (Persaud 2003).

Stress testing a portfolio under a variety of assumptions about correlations between trading costs, asset returns, and the trading induced by the strategy should give a better picture of the liquidity risk in the portfolio.

It seems plausible that many investment managers experience autocorrelated net flows into or out of the fund (e.g., a fund that has done well will tend to receive inflows while a fund that has done poorly is more likely to experience redemptions). Evidence of autocorrelated fund flows and institutional trading is found by Del Guercio and Tkac (2002), Campbell et al. (2009), Frazzini and Lamont (2008), and Lou (2008). With autocorrelated fund flows, following the optimal trading models in Almgren and Chriss (2000), Grinold and Kahn (2000), Huberman and Stanzl (2005), Hora (2006), and Engle and Ferstenberg (2007) may cause predictable price pressure for the assets held in the portfolio. For example, assume that Fund A receives a cash inflow and wants to buy shares of XYZ today using the trading strategy in Almgren and Chriss (2000). If Fund A also receives a fund inflow tomorrow, its traders are likely to want to buy more shares of XYZ, probably still using the Almgren and Chriss (2000) strategy. This implies that the fund will be buying XYZ more aggressively in the morning, both yesterday and today.

Heston et al. (2009) study the intraday patterns in stock returns by estimating cross-sectional regressions in which returns over each half-hour intraday period are regressed on half-hour returns j periods ago, where j runs from 1 to 520. The coefficient is negative for low lags, as one would expect given the fact that bid–ask bounce would induce a negative coefficient. However, at lags that are multiples of 13 (which corresponds to the same half-hour interval on different days) the coefficients are positive and statistically significant out to 520 lags (which corresponds to forty trading days). Thus, whether asset i had a high return in the 1:30 P.M.–2:00 P.M. time slot forty days ago has statistically significant explanatory power for its return in the 1:30 P.M.–2:00 P.M. time slot today. This periodicity does not provide an arbitrage opportunity, given the size of the bid–ask spreads. However, the periodicity might help traders time their trades. While there may be alternative explanations for this empirical regularity, persistence in order flows linked with trading algorithms might be the reason.